Chapter 4

Unconfounded Observational Studies

We have learned that experiments allow for efficient and unbiased causal inference, through controlled (randomized) treatment assignment. However, for many problems experiments would be impossible to implement, or unethical, or too costly. For example we cannot randomize people to smoke, or become obese; an experiment studying lifetime diets on mortality would take decades (and be riddled with noncompliance). Further, detailed observational data is often collected and readily available: what should be done with such data?

In some ways observational studies can resemble experiments. For example the observed data structure is the same – we still observe a triplet of covariates, treatment, and outcome $(X, A, Y) \sim \mathbb{P}$. We will also continue to assume the consistency condition ($Y = Y^a$ if $A = a$), so that the observed outcome equals the potential outcome under the observed treatment.

However there is a major underlying difference: in observational studies, the treatment happened "naturally" according to some unknown process, and was not under experimenters' control. Consider some examples. Cancer patients may decide to have surgery or not based on myriad factors: current health, past medical history, conversations with doctors and family, assessment of risk-benefit trade-offs, etc. Class size is not random and instead depends on popularity of subject matter, quality of lecturer, etc. Gun laws vary widely across US states; they are most restrictive in northeast states, and least restrictive in northwest and southeast states. Reasons for this variation might include, for example, cultural differences or reactions to specific shootings or events (e.g., bans on assault weapons and bump stocks were introduced after shootings in 2012 at Sandy Hook Elementary School and in 2017 in Las Vegas, respectively).

Mathematically, this means that in observational studies the treatment distribution

$$\mathbb{P}(A = a \mid X, Y^{a'})$$

is unknown. In contrast, recall that in Bernoulli experiments it is known by design that $\mathbb{P}(A = 1 \mid X, Y^a) = \mathbb{P}(A = 1) = \pi$.

4.1 No Identification Without Assumptions

When the treatment A and potential outcomes Y^a can be correlated, then there is unfortunately no hope for identification. Intuitively this is because consistency only lets us learn about potential outcomes under treatment among those who were actually treated – those who were not may be arbitrarily different.

Formally we can write the mean potential outcome under treatment as

$$\begin{aligned}\mathbb{E}(Y^1) &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^1 \mid A=1)\mathbb{P}(A=1) \\ &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y \mid A=1)\mathbb{P}(A=1)\end{aligned}$$

where the first equality follows by the law of total probability, and the second by consistency. However in general the observed data distribution $\mathbb{P}$ says nothing about the quantity $\mathbb{E}(Y^1 \mid A=0)$. In general, $\mathbb{E}(Y^1 \mid X, A=0) \neq \mathbb{E}(Y^1 \mid X, A=1)$ since those who take control may be completely different from those who take treatment.

The next proposition shows that, if only relying on consistency, one can merely bound rather than point identify mean potential outcomes; further, these bounds are necessarily imprecise and cannot identify whether the treatment has a non-zero effect.

Proposition 4.1. *Let $(A, Y) \sim \mathbb{P}$ with $\mathbb{P}(Y \in [0,1]) = 1$ and assume consistency so that $Y = Y^a$ if $A = a$. Then*

$$\mathbb{E}\{(2A-1)Y\} - \mathbb{P}(A=1) \leq \mathbb{E}(Y^1 - Y^0) \leq \mathbb{E}\{(2A-1)Y\} + \mathbb{P}(A=0)$$

and these bounds are sharp.

Proof. We have

$$\begin{aligned}\mathbb{E}(Y^1) &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^1 \mid A=1)\mathbb{P}(A=1) \\ &= \mathbb{E}(Y^1 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y \mid A=1)\mathbb{P}(A=1) \\ &\in \Big[\mathbb{E}(Y \mid A=1)\mathbb{P}(A=1),\ \mathbb{P}(A=0) + \mathbb{E}(Y \mid A=1)\mathbb{P}(A=1)\Big] \\ &= \Big[\mathbb{E}(AY),\ \mathbb{P}(A=0) + \mathbb{E}(AY)\Big]\end{aligned}$$

where the first equality follows by the law of total probability, the second by consistency, and the last bounds by the fact that $Y \in [0,1]$. By the same logic

$$\begin{aligned}\mathbb{E}(Y^0) &= \mathbb{E}(Y^0 \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^0 \mid A=1)\mathbb{P}(A=1) \\ &= \mathbb{E}(Y \mid A=0)\mathbb{P}(A=0) + \mathbb{E}(Y^0 \mid A=1)\mathbb{P}(A=1) \\ &\in \Big[\mathbb{E}\{Y(1-A)\},\ \mathbb{E}\{Y(1-A)\} + \mathbb{P}(A=1)\Big]\end{aligned}$$

Taking the difference of the lower and upper bounds (and vice versa) yields the result. The lower bound is attained when $Y^a = 1-a$ among those with $A = 1-a$, and the upper bound when $Y^a = a$ among those with $A = 1-a$. □

Remark 4.1. The assumption that $Y \in [0,1]$ with probability one is immaterial as long as Y has bounded support, since any $Y \in [a,b]$ can always be rescaled as $\frac{Y-a}{b-a} \in [0,1]$.

Note that the length of the above bounds is exactly one, so they must necessarily include zero. This implies that without further assumptions it is impossible to rule out whether a treatment has no effect. This should make you question any claim of assumption-free causal inference, or general test for unmeasured confounding.

4.2 Identification

Without any assumptions beyond consistency, we have seen that is impossible to rule out the possibility of zero treatment effect, even with infinite data, and that the value of the treatment effect can at best be bounded. One way to make progress is to try to collect as many relevant covariates X as possible to be able to explain the treatment process, in the sense that

$$A \perp\!\!\!\perp Y^a \mid X. \tag{4.1}$$

Condition (4.1) goes by several names in the literature: exchangeability, ignorability, or no unmeasured confounding. It means treatment is essentially randomized within levels of the covariates, since it is conditionally independent of potential outcomes. In other words, there can be no remaining unmeasured confounders U that may induce a correlation between the treatment and potential outcomes. Condition (4.1) can be illustrated graphically as in the directed acyclic graph in Figure 4.1.

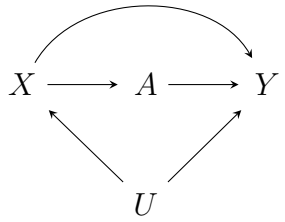


Figure 4.1: A directed acyclic graph representation of the no unmeasured confounding condition (4.1), which holds due to the missing arrow $U \to A$.

Importantly, no unmeasured confounding as in (4.1) means the observational study is actually a conditionally randomized experiment, but one in which the randomization probabilities

$$\mathbb{P}(A = a \mid X = x, Y^{a'} = u) = \mathbb{P}(A = a \mid X = x)$$

are unknown. In other words, an unconfounded observational study can be viewed as a setup where nature ran a conditionally randomized experiment, but kept the randomization probabilities hidden from us.

A critical question is: how does one verify or justify the assumption (4.1)?

Condition (4.1) is generally impossible to test with data, since Y^a is not directly observed (under consistency Y^a is only observed among those with $A = a$, so one cannot measure the correlation between A and Y^a). This means that causal inference cannot be purely data driven – some subject matter knowledge is required. In fact, (4.1) can generally only be justified non-mathematically using subject matter expertise. For example, one might try to understand how physicians are assigning treatment, or how people select into job training programs, etc.

Remark 4.2. The no unmeasured confounding condition (4.1) is commonly invoked, but it is by no means the only assumption or strategy one might consider for causal inference in non-experimental settings. In future chapters we will study alternatives.

Recall that (4.1) also held in the Bernoulli experiments we considered; however there are some important differences. First, in an unconfounded observational study we only have $A \perp\!\!\!\perp Y^a \mid X$, and not $A \perp\!\!\!\perp (X, Y^a)$. This means that there are important differences among subjects receiving different treatment levels (but that these differences are only relevant insofar as they appear in observed covariates). Second, the treatment distribution $\mathbb{P}(A = a \mid X = x)$ is unknown and thus would have to be estimated. Third, we can never really be sure that (4.1) holds in an observational study, whereas in an experiment we know with certainty that it holds due to the design.

Remark 4.3. When treatment is binary, the quantity $\mathbb{P}(A = 1 \mid X = x) = \pi(x)$ is known as the "propensity score". It represents the conditional chance of receiving treatment for subjects with covariates $X = x$.

The strategy for identifying average treatment effects in unconfounded observational studies is essentially the same as that used in experiments. This is formalized in the next proposition.

Proposition 4.2. *Let* $(X, A, Y) \sim \mathbb{P}$ *and assume:*

1. *Consistency:* $Y = Y^a$ *if* $A = a$.
2. *No unmeasured confounding:* $A \perp\!\!\!\perp Y^a \mid X$.
3. *Positivity:* $\mathbb{P}(A = a \mid X = x) > 0$ *with probability one.*

Then

$$\mathbb{E}(Y^a) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}.$$

Proof. We have

$$\begin{aligned}\mathbb{E}(Y^a) &= \mathbb{E}\{\mathbb{E}(Y^a \mid X)\} = \mathbb{E}\{\mathbb{E}(Y^a \mid X, A = a)\} \\ &= \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}\end{aligned}$$

where the first equality follows by iterated expectation, the second by no unmeasured confounding, and the third by consistency. Positivity is required so that the conditional expectations are well-defined. □

In Proposition 4.2 we needed a *positivity* condition, which was implicit in experiments since it held by design. Namely positivity requires that the propensity score be bounded away from extreme values; in other words, to estimate the mean potential outcome if everyone were treated at level $A = a$, we require that everyone has some non-zero chance of being treated at that level. In experiments this holds as long as the proverbial coins that are flipped to decide treatment assignment are not deterministic. However, in observational studies, positivity is a real concern; some subjects may really have zero chance of receiving treatments other than the one they actually received. For example, very sick patients may never not receive treatment, or very healthy patients may never have intensive life-saving surgeries.

Positivity is sometimes called the "experimental treatment assumption" or "overlap". And it can be stated in various more or less equivalent ways. For example, with binary treatments the following are equivalent:

- $0 < \pi(x) < 1$ for all $x \in \mathcal{X}$ with $\mathbb{P}(X = x) > 0$, i.e., $\mathbb{P}\{0 < \pi(X) < 1\} = 1$
- $0 < d\mathbb{P}(x \mid A = 1)/d\mathbb{P}(x \mid A = 0) < \infty$

Positivity is also sometimes written as

- $\mathbb{P}\{\epsilon \leq \pi(X) \leq 1 - \epsilon\} = 1$ for some $\epsilon > 0$

The first two forms are sufficient for identification, while the third is often used for analyzing estimators, since arbitrarily poor performance may be possible if the propensity scores can be arbitrarily close to zero (even if positive). For simplicity, we will often just use the stronger ϵ bound.

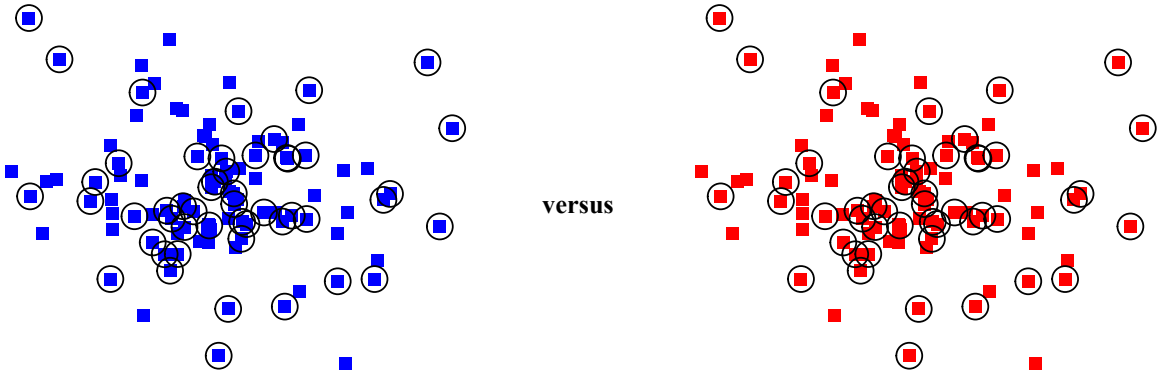
4.3 Effects of Treatment on the Treated

So far we have focused on identification of average treatment effects, e.g., of the form $\mathbb{E}(Y^1 - Y^0)$. However, it is also common (especially in observational studies) to pursue the average treatment effect on the treated given by

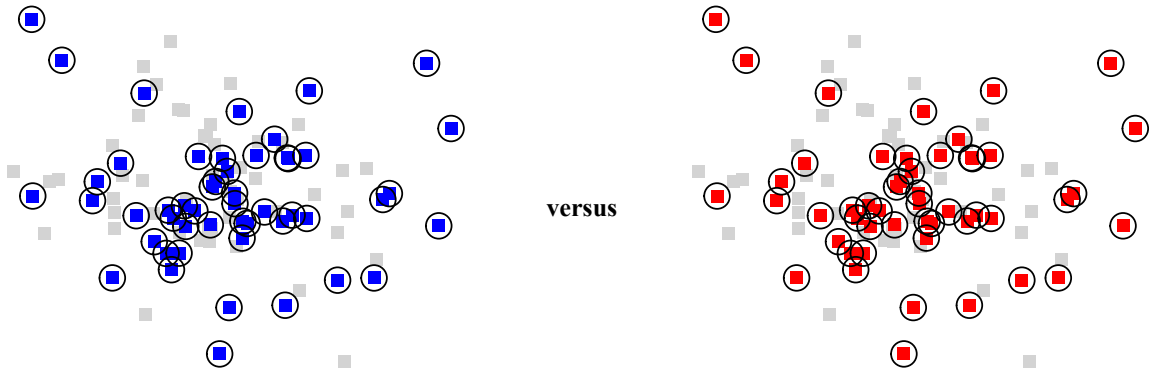
$$\psi_{att} = \mathbb{E}(Y^1 - Y^0 \mid A = 1) = \mathbb{E}(Y - Y^0 \mid A = 1)$$

In contrast to the average treatment effect $\mathbb{E}(Y^1 - Y^0)$ the effect of treatment on the treated measures the difference in mean outcomes if treatment was removed from those who received it. This parameter, and its relation to the average effect, is illustrated in the schematic given in Figure 4.2.

The effect on the treated parameter can be useful if the goal is to learn effects of removing an exposure, e.g., when assigning everyone in the population treatment may not be feasible. It also requires weaker identifying assumptions, as illustrated in the next proposition.



(a) Average treatment effect



(b) Effect on the treated

Figure 4.2: Illustration of average treatment effect versus effect on the treated parameters. Each point in the plot represents a subject in some population. Circled points represent subjects who actually received treatment, and the color of the point represents the treatment they would receive based on the counterfactual contrast of interest, with blue denoting treatment and red control (and gray meaning they are excluded). The average treatment effect in (a) is a comparison of the mean outcome if all subjects were treated (regardless of actual treatment) versus if none were treated. In contrast, the effect on the treated in (b) compares the mean outcome among those actually treated to what it would have been if treatment was removed (among only this subpopulation).

Proposition 4.3. *Let* $(X, A, Y) \sim \mathbb{P}$ *and assume:*

1. *Consistency:* $Y = Y^a$ *if* $A = a$.
2. *No unmeasured confounding:* $A \perp\!\!\!\perp Y^0 \mid X$.
3. *Positivity:* $\mathbb{P}(A = 0 \mid X = x) > 0$ *with probability one among those with* $A = 1$.

Then

$$\mathbb{E}(Y^1 - Y^0 \mid A = 1) = \mathbb{E}\{Y - \mathbb{E}(Y \mid X, A = 0) \mid A = 1\}.$$

Proof. By consistency, it follows that $\mathbb{E}(Y^1 \mid A = 1) = \mathbb{E}(Y \mid A = 1)$. Then

$$\begin{aligned}\mathbb{E}(Y^0 \mid A = 1) &= \mathbb{E}\{\mathbb{E}(Y^0 \mid X, A = 1) \mid A = 1\} \\ &= \mathbb{E}\{\mathbb{E}(Y^0 \mid X, A = 0) \mid A = 1\} \\ &= \mathbb{E}\{\mathbb{E}(Y \mid X, A = 0) \mid A = 1\}\end{aligned}$$

where the first equality follows by iterated expectation, the second by no unmeasured confounding, and the third by consistency. Positivity is required so that the conditional expectations and their averages are well-defined. □

Intuitively, since the first quantity $\mathbb{E}(Y^1 \mid A = 1) = \mathbb{E}(Y \mid A = 1)$ in the effect on the treated is identified under no conditions, one only needs $A \perp\!\!\!\perp Y^0 \mid X$ and similarly "one-sided" positivity, in that the propensity scores only need to be bounded away from one. The latter follows since anyone with $\pi(x) = 0$ will necessarily not be part of the $A = 1$ population, and thus there is no need to assess their outcome had they been treated. On the other hand, if there are subjects with $\pi(x) = 1$ then they would be part of the $A = 1$ population, but learning about their outcome under control would be impossible.

4.4 Observational Studies versus Experiments

The identification results from Propositions 4.2 and 4.3 have converted causal problems into purely statistical ones. For example, in the average treatment effect case, after identification the goal has become to estimate the averaged regression function

$$\mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}$$

as well as possible; this is a purely statistical problem. The difficulty compared to the experimental setup is that in an observational study the treatment distribution is unknown. Before we move to estimation, however, we will discuss the experimental/observational distinction in some more detail.

A fascinating debate can be had about the evidential status of observational studies versus experiments. There are extremists on both sides, but the reality is more nuanced and context-dependent. The main issues stem from validity, feasibility, and generalizability.

Here are some sentiments you might hear from a "Pro-Observational Extremist":

- Causality is easy: just throw basic covariates into a regression model and voila!
- We should not waste time and money on experiments, when we can just do cheap and easy observational studies.

These sentiments are sometimes implicit rather than explicit, but examples abound, for example in exploratory research with catchy titles meant to make headlines.

Here are some sentiments you might hear from a "Pro-Experimental Extremist":

- The only way to learn anything reliable is with the gold standard: experiments.
- All experiments are necessarily trustworthy and informative, and anything else is not real science.

Examples of this kind of extremism can be found among defenses of tobacco and pharmaceutical companies. For example Michaels [2008] recounts a cigarette executive stating that "Doubt is our product, since it is the best means of competing with the 'body of fact' that exists in the minds of the general public." Statisticians are probably more likely to be pro-experimental extremists than pro-observational extremists.

As is often the case with extremist perspectives, real life is more nuanced.

Here are some issues with the pro-observational extremist perspective:

- Causal claims from observational studies require untestable assumptions that can be difficult to assess.
- Adjusting for measured covariates may not be enough, since there could always be some unmeasured confounding that was missed.
- Even if all relevant confounders happened to be measured, appropriate adjustment and valid inference is non-trivial in nonparametric or high-dimensional models.

Here are some issues with the pro-experimental extremist perspective:

- Experiments are not always feasible (e.g.,studies of long-term effects) or ethical (e.g., smoking).
- Experiments are often conducted in selected, non-representative populations: an unbiased estimate in the wrong population may be less useful than a slightly biased estimate in the right one.
- Experiments are often plagued with non-compliance and missing data, which can essentially turn them into observational studies.

In sum, observational studies and experiments are not necessarily uniformly bad or good; both types of studies range in quality, and it is not true that one type dominates the other. Their evidential status is context-dependent and needs to be evaluated case-by-case, based on specific merits or faults.

Appendix A

Notation Guide

Y^a	Potential outcome under treatment/exposure $A = a$
$\perp\!\!\!\perp$	Statistically independent
$\xrightarrow{p}$	Convergence in probability
$\rightsquigarrow$	Convergence in distribution
$O_{\mathbb{P}}(1)$	Bounded in probability
$o_{\mathbb{P}}(1)$	Converging in probability to zero
$\mathbb{P}_n$	Sample average operator, as in $\mathbb{P}_n(\widehat{f}) = \mathbb{P}_n\{\widehat{f}(Z)\} = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}(Z_i)$
$\mathbb{P}$	Conditional expectation given the sample operator, as in $\mathbb{P}(\widehat{f}) = \int \widehat{f}(z)\, d\mathbb{P}(z)$
$\|\cdot\|$	$L_2(\mathbb{P})$ norm $\|f\| = \sqrt{\mathbb{P}(f^2)}$ or Euclidean norm, depending on context
$\|\cdot\|_1$	$L_1(\mathbb{P})$ norm $\|f\|_1 = \mathbb{P}(\|f\|)$
$\|\cdot\|_\infty$	L_∞ or supremum norm $\|f\|_\infty = \sup_z \|f(z)\|$
$\mathcal{H}(s)$	Hölder class of functions with smoothness index s
$\lesssim$	Less than or equal, up to a constant multiplier

Bibliography

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20(3):261, 2005.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

S. Leon, A. A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.

D. Michaels. *Doubt is their product: how industry's assault on science threatens your health*. Oxford University Press, 2008.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer, 1982.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1): 113–124, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.